Discussion of "Estimation and Accuracy after Model Selection" by Brad Efron

Lan Wang, Ben Sherwood and Runze Li

We congratulate Efron for his stimulating and timely work which addresses an important issue on estimation after model selection. In practice, it is typical to ignore the variability of the variable selection step, which could result in inaccurate post-selection inference. Although the flaw of such practice is widely recognized, finding a general solution is extremely challenging. The model selection step is often a complex decision process and can involve collecting expert opinions, preprocessing, applying a variable selection rule, data-driven choice of one or more tuning parameters, among others. Except in simple cases, explicitly characterizing the form of the post-selection estimator is itself difficult. The key result of this paper is a closed-form formula for obtaining the standard deviation of a "*bootstrap smoothed*" (or "*bagged*") estimator. This elegant formula is not only simple to implement but also versatile. It indeed provides a general approach for obtaining a confidence interval for a class of parameters of interest while incorporating the variability of variable selection.

Our discussions will focus on two aspects: (1) the generality of the method, and (2)

[1]Lan Wang is Associate Professor and Ben Sherwood is graduate student, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Email: wangx346@umn.edu. Runze Li is Distinguished Professor, Department of Statistics and the Methodology Center, the Pennsylvania State University, University Park, PA 16802-2111. Email: rzli@psu.edu. Wang and Sherwood's research is supported by a NSF grant DMS1308960. Li's research is supported by NIDA, NIH grants P50 DA10075 and P50 DA036107. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

further insight into the performance of the proposed method in a simple but hopefully informative example.

1 Generality of the method

In principle, the standard deviation formula in Efron's Theorem 1 can be applied to general "*bootstrap smoothed*" (or "*bagged*") estimators. As the central example of the paper is traditional linear regression, we empirically investigate the performance of the proposed estimator in a variety of regression settings where the proposed method is expected to be useful through Monte Carlo simulations. In particular, we will consider: (1) LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001) for linear regression, (2) Poisson regression as a representative example of generalized linear models, (3) quantile regression for predicting a conditional quantile, and (4) nonparametric regression where we apply a data-driven method to select the number of spline basis functions (this last example was motivated by a discussion with Professor Xuming He).

For each of the four cases, we construct confidence intervals for the conditional mean (or quantile) using the new method proposed in Efron's paper (denoted by "new"). We compare the new method with the standard bootstrap confidence interval (denoted by "standard") and the percentile interval (denoted by "percentile"), as described in Efron's paper.

1.1 Several numerical examples

Example 1. (Regularized estimators for linear regression) The response variable is generated from the model $Y = 1 + X_1 - X_3 + X_6 + \epsilon_i$, where the candidate covariates $X_1, \ldots, X_6$ are independent standard normal random variables. The ran-

dom error ϵ is normally distributed with mean zero and standard deviation 2, and is independent of the covariates. The sample size is $n = 200$. The main goal is to study the proposed method when regularized methods such as LASSO and SCAD are used to obtain the selected model. We implement LASSO using the R package glmnet and implement SCAD using the coordinate descent algorithm in the R package ncvreg. For LASSO, we use five-fold cross validation to select the tuning parameter; while for SCAD we apply BIC (Wang, Li and Tsai, 2007) for selecting the tuning parameter. For completeness, we also include best subset selection procedures based on C_p, AIC and BIC.

We consider the 95% confidence interval for estimating the conditional mean at $X = (-2.5, -2.5, -2.5, -2.5, -2.5, -2.5)'$. The results are summarized in Table 1 based on 4000 bootstrap samples. We assess the performance by the length of the confidence interval and its coverage probability (reported in the last two columns of the table). The third column reports the center of the confidence interval.

Table 1: Linear Regression

Method	Interval Type	Center	Length	Coverage
C_p	new	-1.58	3.25	0.98
	percentile	-1.58	3.46	0.98
	standard	-1.54	3.46	1.00
AIC	new	-1.58	3.26	0.97
	percentile	-1.58	3.46	0.98
	standard	-1.54	3.46	1.00
BIC	new	-1.56	2.93	0.98
	percentile	-1.56	3.23	0.98
	standard	-1.52	3.22	0.99
LASSO	new	-1.47	3.33	0.96
	percentile	-1.49	3.39	0.97
	standard	-1.40	3.38	0.95
SCAD	new	-1.55	2.97	0.98
	percentile	-1.55	3.25	0.98
	standard	-1.51	3.24	0.99

Example 2. (Poisson regression) The response variable is generated from the

model $Y \mid X \sim \text{Poisson}\left(e^{1+X-X^2}\right)$ where X has a standard normal distribution. The sample size is $n = 400$. We use AIC and BIC for model selection. For candidate models, we consider different polynomial degrees of X, from linear to sextic. The results for the confidence interval for estimating $E[Y \mid X = -2]$ are reported in Table 2 based on 6000 bootstrap runs.

Table 2: Poisson Regression

Method	Interval Type	Center	Length	Coverage
AIC	new	20.13	3.09	0.97
	percentile	20.14	3.75	0.99
	standard	20.18	3.71	0.99
BIC	new	20.12	2.07	0.97
	percentile	20.13	2.56	0.97
	standard	20.11	2.54	0.98

Example 3. (Quantile regression) The response variable is generated from the heteroscedastic regression model $Y = 1 + 3X_1 - 1.5X_3 + 2X_6 + (1 + X_2)\epsilon$, where the X_i's, $i = 1, \ldots, 6$, are independent and uniformly distributed on $(0, 1)$. The random error ϵ has a standard normal distribution and is independent of the X_i's. The sample size is $n = 200$.

We considered AIC and BIC for model selection, which are based on the quantile loss function and programmed in the *quantreg* package in R. Penalized quantile regression with LASSO or SCAD penalty is also considered. The results for the confidence interval for estimating the 0.7 conditional quantile at $X = (0.9, \ldots, 0.9)'$ are reported in Table 3 based on 4000 bootstrap runs.

Example 4. (Nonparametric regression) The response variable is generated from the regression model $Y = 1 + X^2\exp(X) + \epsilon$, where X is uniformly distributed on $(0, 1)$. The random error ϵ is normally distributed with mean zero and standard deviation 2, and is independent of X. The sample size is $n = 100$.

We estimate the nonparametric regression function via B-spline regression. We

Table 3: Quantile regression

Method	Interval Type	Center	Length	Coverage
AIC	new	5.10	1.85	0.95
	percentile	5.11	2.12	0.98
	standard	5.08	2.12	1.00
BIC	new	5.07	1.77	0.94
	percentile	5.09	2.12	0.97
	standard	5.05	2.13	0.97
LASSO	new	5.00	1.73	0.94
	percentile	5.02	2.00	0.95
	standard	5.03	2.02	0.96
SCAD	new	5.06	1.77	0.94
	percentile	5.08	2.11	0.98
	standard	5.05	2.12	0.97

select the number of knots (ranging from 1 to 5) by a BIC criterion. More specifically, let ν represent the number of degrees of freedom of a candidate model and let $\hat{\sigma}_\nu^2$ be the estimate of σ^2 for the corresponding model. We select the model that minimizes $\mathrm{BIC}(\nu) = n\log(\hat{\sigma}_\nu^2) + \nu\log(n)$, see, for example, He and Shi (1996). The results for the confidence interval for estimating the conditional mean at $X = 0.9$ are reported in Table 4 based on 4000 bootstrap runs.

Table 4: Nonparametric Regression

Interval Type	Center	Length	Coverage
new	2.99	1.83	0.93
percentile	2.99	2.18	0.97
standard	2.99	2.16	0.97

1.2 Observations from the numerical examples

In the above examples, we observe that the new confidence interval proposed in Efron's paper provides a more accurate confidence interval for all cases and keeps better coverage rates for most cases than the standard interval and the percentile interval when the estimator is obtained after variable selection.

From our limited simulation experience, we note that the choice of the number of

bootstrap samples is important to the performance of the new method. A suitable choice of B can vary depending on the underlying model and the amount of noise in the data. We find that $B = 4000$ works reasonably well for most of the situations we have considered.

An interesting observation from our simulations is that the new method can also be useful for regularized procedures, in particular SCAD, when the tuning parameter is chosen in a data-driven fashion. It is known that the "*bootstrap smoothed*" (or "*bagged*") estimators are most valuable when hard decisions rules (such as best subset selection, decision trees) are involved, which result in instability in prediction. In practice, when a regularization procedure such as LASSO or SCAD is applied, the tuning parameter is often selected by cross-validation or a modified BIC, which introduces extra variability in the final estimator. Although the improvement over Lasso is sometimes marginal as Efron has pointed out, it may still be worthwhile (in the quantile regression example, we observe a 15% reduction of interval length for Lasso). For SCAD, with the tuning parameter being selected by BIC, the improvement is more significant. Our simulation experience, including that not reported here due to space limitation, indicates that the gain of the new method is more pronounced when the sample size is smaller and the data are noisier.

2 Further insight from a simple example

Next, we will consider Efron's main example in the orthogonal regression case, which sheds some light on its performance. Let Y be the $n \times 1$ vector of responses and $X = (X_1, \ldots, X_p)^T$ be the design matrix. It is assumed that $X^T X = nI_n$, where I_n is the $n \times n$ identity matrix. The least squares estimator for β_j is $\widehat{\beta}_j = n^{-1} X_j^T Y$.

For a given model M, where M denotes an index set for the covariates in the model, Mallow's C_p is defined as $C_p(M) = (Y - X_M\widehat{\beta}_M)^T(Y - X_M\widehat{\beta}_M) + 2\sigma^2|M|$, where X_M denotes the submatrix of X corresponding to M, and β_M denotes the least squares estimator for model M. In the orthogonal regression case, it is easy to see

$$C_p(M) \;\; = \;\; Y^TY + \sum_{j\in M}(-n\widehat{\beta}_j^2 + 2\sigma^2).$$

As a result, C_p selects all X_j such that $-n\widehat{\beta}_j^2 + 2\sigma^2 < 0$. Hence, given a vector of covariates $x = (x^{(1)}, \ldots, x^{(p)})$, the estimator of $E(Y|X = x)$ obtained after applying Mallow's C_p criterion can be written as

$$\sum_{j=1}^{p} x^{(j)}\widehat{\beta}_j I(|\widehat{\beta}_j| > \sigma\sqrt{2/n}).$$

Since the effect of each covariate is additive, we consider the univariate case in the following discussion. The post-selection estimator of the conditional mean at x is,

$$t_n(Y|x) = x\widehat{\beta}I(|\widehat{\beta}| > \sigma\sqrt{2/n}).$$

The bootstrap smoothed estimator given by Efron is

$$s_n(Y|x) = B^{-1}\sum_{i=1}^{B} t_n(Y^*|x),$$

where Y^* is the bootstrap sample.

The asymptotic distribution of $s_n(Y|x)$ is known under a local asymptotic framework. Assume that $Y_i = \beta X_i + \epsilon_i$, where $\beta = \beta_n(b) = b\sigma n^{-1/2}$ for some constant b, $X_1, \ldots, X_n$ are i.i.d. random variables with $\mathrm{E}(X_i^2) = 1$, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. and

independent from the X_i's, $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2 < \infty$. It follows Proposition 2.2 of Buhlmann and Yu (2002) that,

$$n^{1/2}\sigma^{-1}s_n(Y|x) \to g_B(Z_b|x)$$

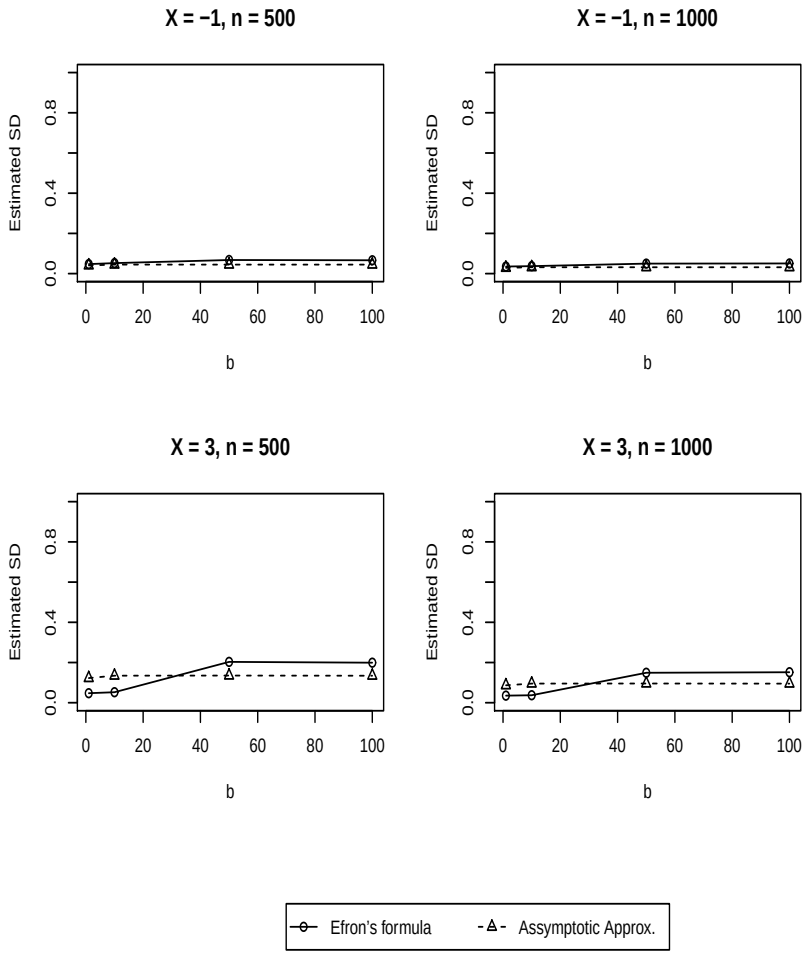
in distribution, where $Z_b = b + Z$, $Z \sim N(0,1)$, and $g_B(z|x) = \big(z - \{z\Phi(\sqrt{2} - z) - \phi(\sqrt{2} - z) - z\Phi(-\sqrt{2} - z) + \phi(-\sqrt{2} - z)\}\big)x$, with Φ and ϕ denoting the distribution function and density function of the standard normal distribution, respectively. The theory thus suggests that the bootstrap smoothed estimator has approximate standard deviation $n^{-1/2}\sigma \times \mathrm{sd}(g_B(z|x))$, where $\mathrm{sd}(g_B(z|x))$ denotes the standard deviation of the distribution given by $g_B(z|x)$.

In Figure 1, we compare the estimated standard deviation of $s_n(Y|x)$ using Efron's formula with that obtained from the above asymptotic distribution (based on simulating the distribution of $g_B(z|x)$) for different values of b at $x = -1$ and 3, for sample sizes $n = 500$ and 1000. The two curves are quite close to each other, suggesting that Efron's estimator performs well in this setting. It is noted that AIC and BIC can be analyzed similarly in the orthogonal design case.

3 Conclusions

Two intriguing questions on Efron's new procedure is: (1) Is it possible to derive the asymptotic property, such as consistency? (2) Can the nonparametric delta method used for deriving the standard deviation formula be extended to the case the number of covariates p_n grows with n? Positive answers to these questions will greatly extend the scope of the application of the new method.

As the bootstrap smoothed estimator combines estimators from different candidate


X = −1, n = 500
X = −1, n = 1000
X = 3, n = 500
X = 3, n = 1000
Estimated SD
0.0
0.4
0.8
0
20
40
60
80
100
b
Efron's formula
Assymptotic Approx.

models, it is may be applicable to situations where we would like to seek inference for a particular parameter of one selected model, unless such a parameter is common to all models. However, we demonstrated that Efron's estimator is useful in a variety of settings when prediction is the goal. Even for a "soft" procedure such as LASSO or SCAD, it can sometimes have notable improvement over existing procedures, when the tuning parameter of such a procedure is selected by a data-driven method.

We greatly appreciate the opportunity of discussing this stimulating work and congratulate the author for his important contributions to this challenging problem.

References

[1] Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, **30**, 927-961.

[2] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

[3] He, X., and Shi, P. (1996) Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis*, **58**, 162181.

[4] Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.